

Метрики и базовые подходы

Петр Погорелов

26 сентября 2024 г.



Еще раз об измерении качества идей
ооооо

Оффлайн эксперимент
оооо

Релевантность
оооооооооо

Покрытие
ооо

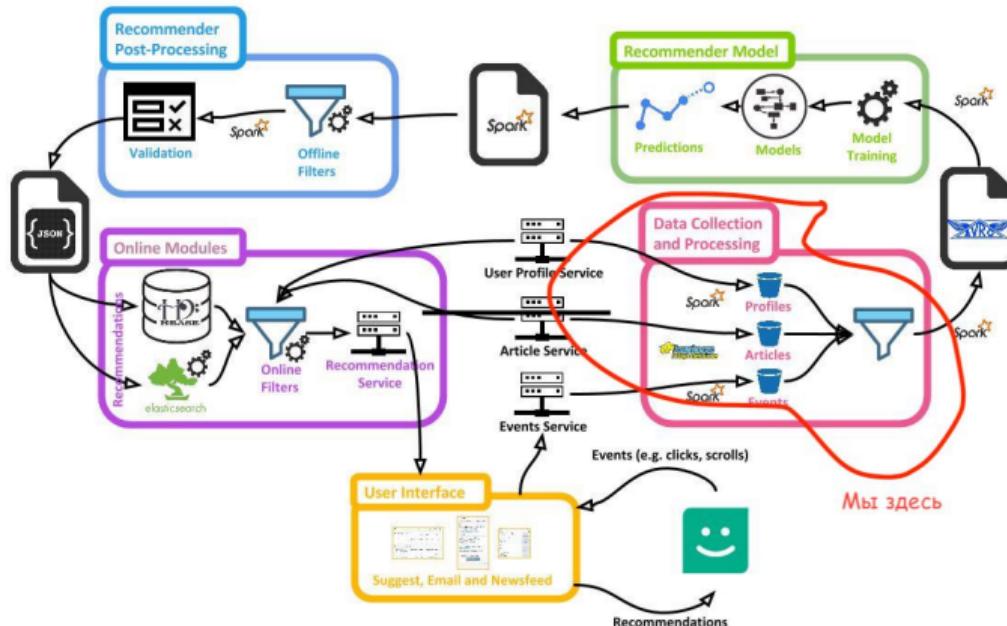
Разнообразие
оооо

Удачность
ооооооо

Бейзлайны
оо

Итоги
оооо

Контекст



Еще раз об измерении качества идей

●oooo

Оффлайн эксперимент
oooo

Релевантность
oooooooo

Покрытие
ooo

Разнообразие
oooo

Удачность
ooooooo

Бейзлайны
oo

Итоги
oooo

Еще раз об измерении качества идей



Миссия компании

Дать пользователям возможность быть ближе к семье и друзьям



Q: Какую метрику вы бы предложили измерять в A/B?

A/B эксперимент [RRSK10]

Плюсы

- Надежная оценка эффекта на любую метрику

Минусы

- Риск необратимо расстроить пользователей
- Дорого заводить
- Долго ждать результат
- Метрик не всегда достаточно

надежность ★ ★ ★

гибкость ★ ★ ★

скорость ★ ★ ★



Опрос пользователей

Плюсы

- Полный контроль над экспериментом
- Оценка эффекта на любую метрику
- Собрать фидбэк напрямую

Минусы

- Дорогой сбор данных
- Смещение аудитории
- Нечестный фидбэк

надежность ★ ★ ★

гибкость ★ ★ ★

скорость ★ ★ ★



Оффлайн эксперимент

Плюсы

- Большая скорость проверки гипотез
- Нельзя сломать прод

Минусы

- Не все метрики доступны оффлайн
- Смещение выборки
- Результат не обязан обобщаться

надежность ★ ★ ★

гибкость ★ ★ ★

скорость ★ ★ ★



Еще раз об измерении качества идей
ооооо

Оффлайн эксперимент
●ооо

Релевантность
ooooooooo

Покрытие
ooo

Разнообразие
oooo

Удачность
ooooooo

Бейзлайны
oo

Итоги
оооо

Оффлайн эксперимент



Какие бывают метрики

Бизнесовая

напрямую интересует бизнес

- сложно оптимизировать
- сложно понять, как компоненты системы влияют на метрику
- сложно мерить офлайн

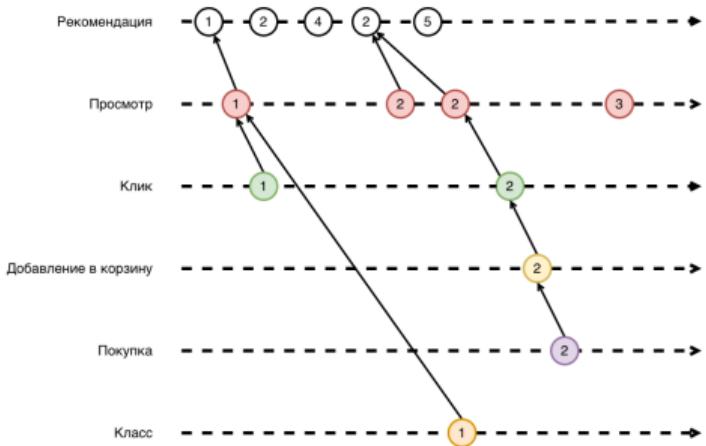
Техническая

отражает один аспект системы

- можно оптимизировать
- можно померить офлайн
- не интересует бизнес :(



Как выбрать техническую метрику



Фидбэк на рекомендации

- Явный/explicit
- Неявный/implicit
- Отложенный/delayed

Какие пользовательские данные использовать для метрики?

- События, интересные бизнесу: их мало и долго ждать
- Быстрые события: их много, но они хуже отражают задачу

- Для измерения качества идей используются все три подхода, но офлайн эксперимент особенно удобен в рекомендательных задачах.
- При подготовке офлайн эксперимента нужно выбрать метрики, которые будут отражать бизнес задачу, и при этом удобно вычисляться.
- Нужно правильно организовать тестовую выборку, например исключить data leak во времени.

Еще раз об измерении качества идей
ооооо

Оффлайн эксперимент
оооо

Релевантность
●ooooooooo

Покрытие
ooo

Разнообразие
ооооо

Удачность
ооооооо

Бейзлайны
oo

Итоги
оооо

Релевантность



Релевантность

Выберите из списка три лучших на ваш вкус фильма



Еще раз об измерении качества идея
ооооо

Оффлайн эксперимент
оооо

Релевантность
оо●оооооо

Покрытие
ооо

Разнообразие
оооо

Удачность
ооооооо

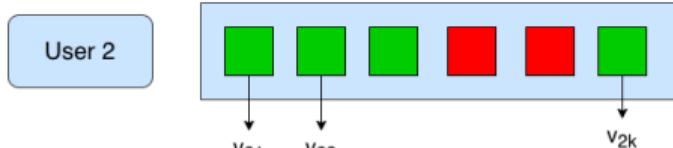
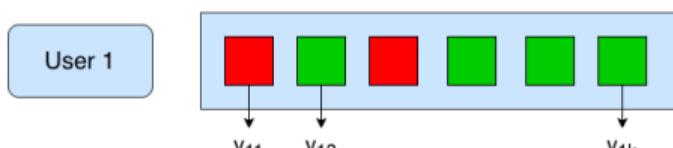
Бейзлайны
оо

Итоги
оооо

Метрики точности

 Non-relevant item

 Relevant item



RMSE, MAE, accuracy, precision, recall, auc, ...



Еще раз об измерении качества идея
ооооо

Оффлайн эксперимент
оооо

Релевантность
ооо●ооооо

Покрытие
ооо

Разнообразие
оооо

Удачность
ооооооо

Бейзлайны
оо

Итоги
оооо

Метрики ранжирования



Non-relevant item

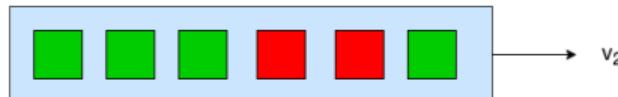


Relevant item

User 1

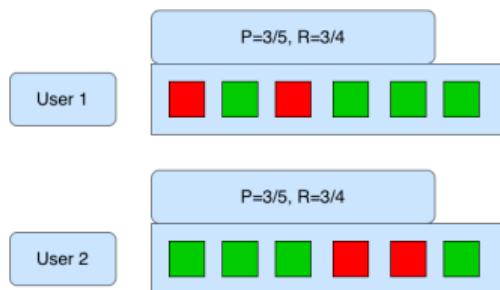


User 2



Precision@k, Recall@k

- Non-relevant item
- Relevant item

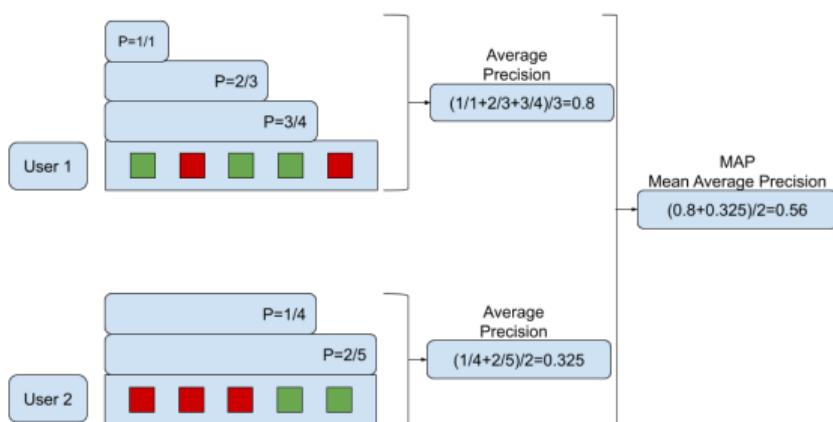


- Легко интерпретировать
- Легко реализовать

- Нечувствительны к порядку внутри k
- Не дают общей картины для любого k

Mean Average Precision [Tai19]

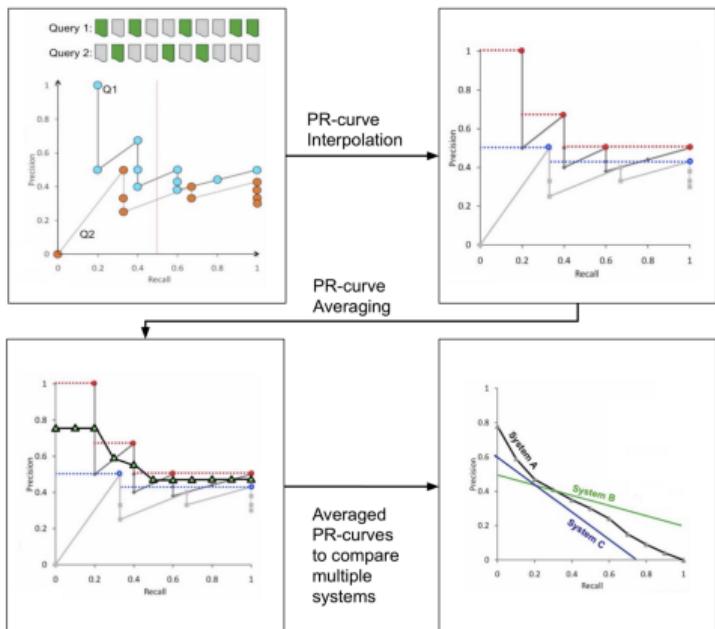
■ Relevant Item
■ Non-Relevant Item



- Дают общую картину качества
- Больше внимания айтемам в голове списка

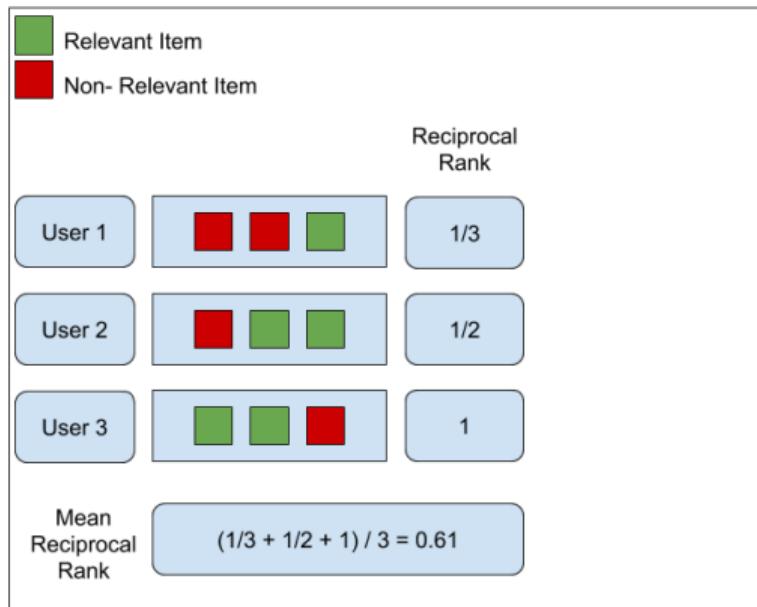
- Подходит только для бинарного фидбэка

Area Under Precision-Recall curve



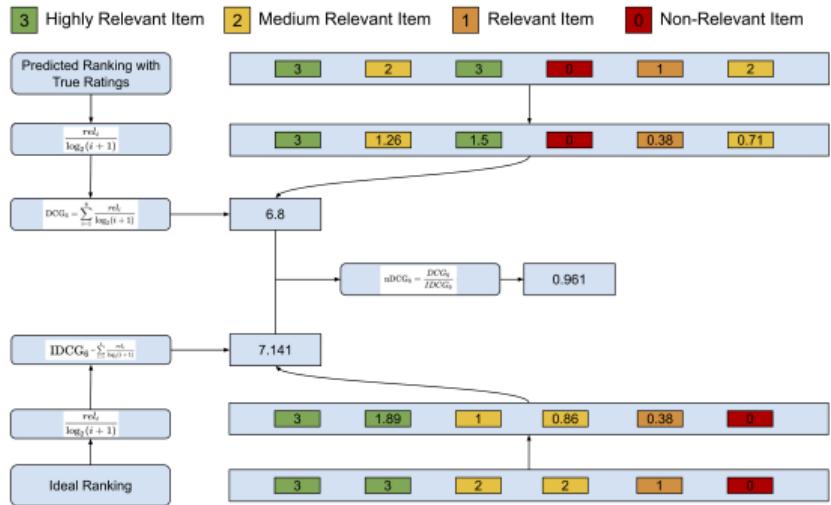
Визуальное представление
MAP

MRR



- Легко интерпретировать
- Легко реализовать
- Удобна для задач, где имеет значение первый результат

- Учитывает только первый результат
- Быстро убывает



- Учитывает не только бинарный фидбэк
- Хорошо учитывает позицию

- Сложно интерпретировать

Еще раз об измерении качества идей
ооооо

Оффлайн эксперимент
оооо

Релевантность
ooooooooo

Покрытие
●oo

Разнообразие
ооооо

Удачность
ooooooo

Бейзлайны
oo



Покрытие

Еще раз об измерении качества идея
ооооо

Оффлайн эксперимент
оооо

Релевантность
ooooooooo

Покрытие
oo•o

Разнообразие
ооооо

Удачность
ooooooo

Бейзлайны
oo

Итоги
оооо

Item space coverage

Какую долю из всех возможных айтемов умеет рекомендовать сервис?

$$cov = \frac{|I_p|}{|I|}$$

$$gini = \frac{1}{|I|-1} \sum_{j=1}^{|I|} (2j - |I| - 1)p(I_j)$$

$p^1(I_j)$ – частота, с которой пользователи выбирают айтем I_j

$p^2(I_j)$ – частота, с которой рекомендер показывает айтем I_j

Айтемы отсортированы по возрастанию $p(I_j)$



Еще раз об измерении качества идей
ооооо

Оффлайн эксперимент
оооо

Релевантность
ooooooooo

Покрытие
oo●

Разнообразие
ооооо

Удачность
ooooooo

Бейзлайны
оо

Итоги
оооо

User space coverage

Доля пользователей, которые могут получить рекомендации



Еще раз об измерении качества идей
ооооо

Оффлайн эксперимент
оооо

Релевантность
ooooooooo

Покрытие
ooo

Разнообразие
●ooo

Удачность
ooooooo

Бейзлайны
oo

Итоги
оооо

Разнообразие



Разнообразие [KP17]

[diversity] Насколько разнообразные айтемы в списке рекомендаций пользователя?



Еще раз об измерении качества идей
ooooo

Оффлайн эксперимент
oooo

Релевантность
oooooooo

Покрытие
ooo

Разнообразие
oo●o

Удачность
oooooo

Бейзлайны
oo

Итоги
oooo

$$div(u) = \frac{\sum_{i=1}^n \sum_{j=1}^n (1 - similarity(i,j))}{n/2(n-1)}$$

With 1% precision loss, percentage of rec. long-tail items increases from 16 to 32, with 5% loss perc. increases to 58.

Метрика сильно зависит от того, как определить сходство



Еще раз об измерении качества идей
ооооо

Оффлайн эксперимент
оооо

Релевантность
ooooooooo

Покрытие
ooo

Разнообразие
ooo●

Удачность
ooooooo

Бейзлайны
oo

Итоги
оооо

Maximal Marginal Relevance [CG98]

$$MMR = \max_j \left[\lambda \text{similarity}(j, U) - (1 - \lambda) \max_{k < j} \text{similarity}(k, j) \right]$$



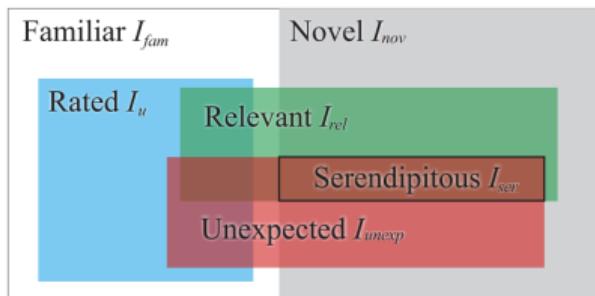
Еще раз об измерении качества идей	Оффлайн эксперимент	Релевантность	Покрытие	Разнообразие	Удачность	Бейзлайны	Итоги
ооооо	оооо	ooooooooo	ooo	оооо	●oooooooo	оо	ооооо

Удачность



Удачность

The term **serendipity** has been recognized as one of the most untranslatable words. The first known use of the term was found in a letter by Horace Walpole to Sir Horace Mann on January 28, 1754. The author described his discovery by referencing a Persian fairy tale, “The Three Princes of Serendip”. The story described a journey taken by three princes of the country Serendip to explore the world. In the letter, Horace Walpole indicated that the princes were “always making discoveries, by accidents and sagacity, of things which they were not in quest of”. [KWV16]



Еще раз об измерении качества идей
ооооо

Оффлайн эксперимент
оооо

Релевантность
оооооооооо

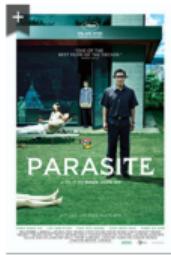
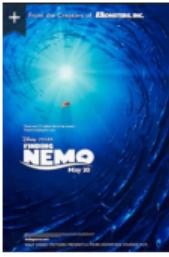
Покрытие
ооо

Разнообразие
оооо

Удачность
оо●оооо

Бейзлайны
оо

Итоги
оооо



Еще раз об измерении качества идей
ооооо

Оффлайн эксперимент
оооо

Релевантность
ooooooooo

Покрытие
ooo

Разнообразие
оооо

Удачность
ooo●ooo

Бейзлайны
oo

Итоги
оооо

Новизна

[novelty] Насколько айтем неизвестен пользователю?

Идея 1: Насколько айтемы близки к айтэмам из истории пользователя?

$$nov^1(u, i) = \min_{j \in I_u} dist(j, i)$$

Идея 2: Насколько айтэмы близки к популярным?

$$nov^2(u, i) = 1 - \frac{|U_i|}{|U|}$$



Еще раз об измерении качества идей
ооооо

Оффлайн эксперимент
оооо

Релевантность
ooooooooo

Покрытие
ooo

Разнообразие
оооо

Удачность
oooo●oo

Бейзлайны
oo

Итоги
оооо

Неожиданность

[unexpectedness] Насколько пользователь ожидает увидеть в рекомендациях айтем?

$$nPMI(i, j) = -\log \frac{p(i, j)}{p(i)p(j)} / \log p(i, j)$$

$$unexp(u, i) = \max_{j \in I_u} (-nPMI(i, j))$$



Еще раз об измерении качества идей	Оффлайн эксперимент	Релевантность	Покрытие	Разнообразие	Удачность	Бейзлайны	Итоги
ооооо	оооо	ooooooooo	ooo	ооооо	ооооо●о	оо	ооооо

Цель	rel	cov	div	ser	poll
Бизнесу					
Увеличить продажи	✓			✓	
Продвигать более разнообразные айтемы		✓	✓		
Улучшить пользовательский опыт	✓		✓	✓	✓
Добиться большей лояльности					✓
Лучше понимать пользователей					✓
Пользователям					
Найти лучший товар	✓			✓	✓
Найти все подходящие товары	✓	✓			✓
Найти последовательность или набор товаров	✓			✓	✓
Залипнуть	✓		✓	✓	✓



Еще раз об измерении качества идей
ооооо

Оффлайн эксперимент
оооо

Релевантность
ooooooooo

Покрытие
ooo

Разнообразие
оооо

Удачность
oooooo●

Бейзлайны
оо

Итоги
оооо

- В оффлайн эксперименте выбираем метрики, отражающие важные аспекты задачи.
- Сперва делаем максимально просто – все равно что-то пойдет не так, и метрики придется допиливать.



Еще раз об измерении качества идей
ооооо

Оффлайн эксперимент
оооо

Релевантность
ooooooooo

Покрытие
ooo

Разнообразие
ооооо

Удачность
ооооооо

Бейзлайны
●○

Итоги
оооо

Бейзлайны



Еще раз об измерении качества идей
ооооо

Оффлайн эксперимент
оооо

Релевантность
ooooooooo

Покрытие
ooo

Разнообразие
оооо

Удачность
ооооооо

Бейзлайны
о●

Итоги
оооо

Простые бейзлайны

- позволяют определить нижнюю границу качества системы
- позволяют быстро стартануть

- Живительный рандом
- TopPopular
- Эвристики
- Редакторская подборка



Еще раз об измерении качества идей
ооооо

Оффлайн эксперимент
оооо

Релевантность
ooooooooo

Покрытие
ooo

Разнообразие
ооооо

Удачность
ooooooo

Бейзлайны
oo

Итоги
●ooo

Итоги



Еще раз об измерении качества идей
ооооо

Оффлайн эксперимент
оооо

Релевантность
ooooooooo

Покрытие
ooo

Разнообразие
ооооо

Удачность
ооооооо

Бейзлайны
oo

Итоги
о•оо

Итоги

При выборе подхода к проверке гипотез, нужно иметь в виду компромисс надежности и скорости

Технические метрики отражают разные аспекты рекомендаций: релевантность, разнообразие, удачность

Don't be a hero: не связываемся со сложными алгоритмами, пока не заведем простые бейзлайны



Еще раз об измерении качества идей
ооооо

Оффлайн эксперимент
оооо

Релевантность
ooooooooo

Покрытие
ooo

Разнообразие
ооооо

Удачность
ооооооо

Бейзлайны
oo

Итоги
оо●○

Подпишис



збазибо!

<https://t.me/mlvok>



Литература I

-  Jaime G. Carbonell and Jade Goldstein, *The use of MMR, diversity-based reranking for reordering documents and producing summaries*, Research and Development in Information Retrieval, 1998, pp. 335–336.
-  Matevz Kunaver and Tomaz Pozrل, *Diversity in recommender systems - a survey*, Knowl. Based Syst. 123 (2017), 154–162.
-  Denis Kotkov, Shuaiqiang Wang, and Jari Veijalainen, *A survey of serendipity in recommender systems*, Knowledge-Based Systems 111 (2016).
-  Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, *Recommender systems handbook*, 1st ed., Springer-Verlag, Berlin, Heidelberg, 2010.
-  Moussa Taifi, *Mrr vs map vs ndcg: Rank-aware evaluation metrics and when to use them*, Nov 2019.

